Toward Millisecond IGP Convergence

Cengiz Alaettinoglu
Van Jacobson
Haobo Yu

`{cengiz,van,haoboy}@packetdesign.com`

Packet Design, Inc.

October 24, 2000

NANOG 20
October 22–24, 2000
Washington, D.C.

Packet Design

Sub-second re-route times give:

- increased network reliability
- support for multi-service traffic (e.g., VoIP)
- lower cost/complexity compared to layer 2 protection schemes like SONET.

Where are we today?

Current IP re-route times are typically tens of seconds.

We need to do better. There are two choices:

- ➤ Replace IP routing with something else, e.g., MPLS Fast Failure Recovery.
- ➤ Figure out what’s wrong with IP routing and fix it.

Since improvement usually requires understanding, we think the second choice is prudent.

Where does the time go?

There are three steps to IS–IS or OSPF re-routing:

(1) Detect a local topology change (link up/down or peer reachability).

(2) Flood a Link State Packet (LSP) to tell adjacent peers about the change.

(3) Compute new routes by running a Dijkstra Shortest Path First (SPF) algorithm on the changed topology.

We did experiments to characterize each of these steps for IS–IS running on Cisco 7200s running 12.0S and 12.1P and Juniper M40s running 4.1.

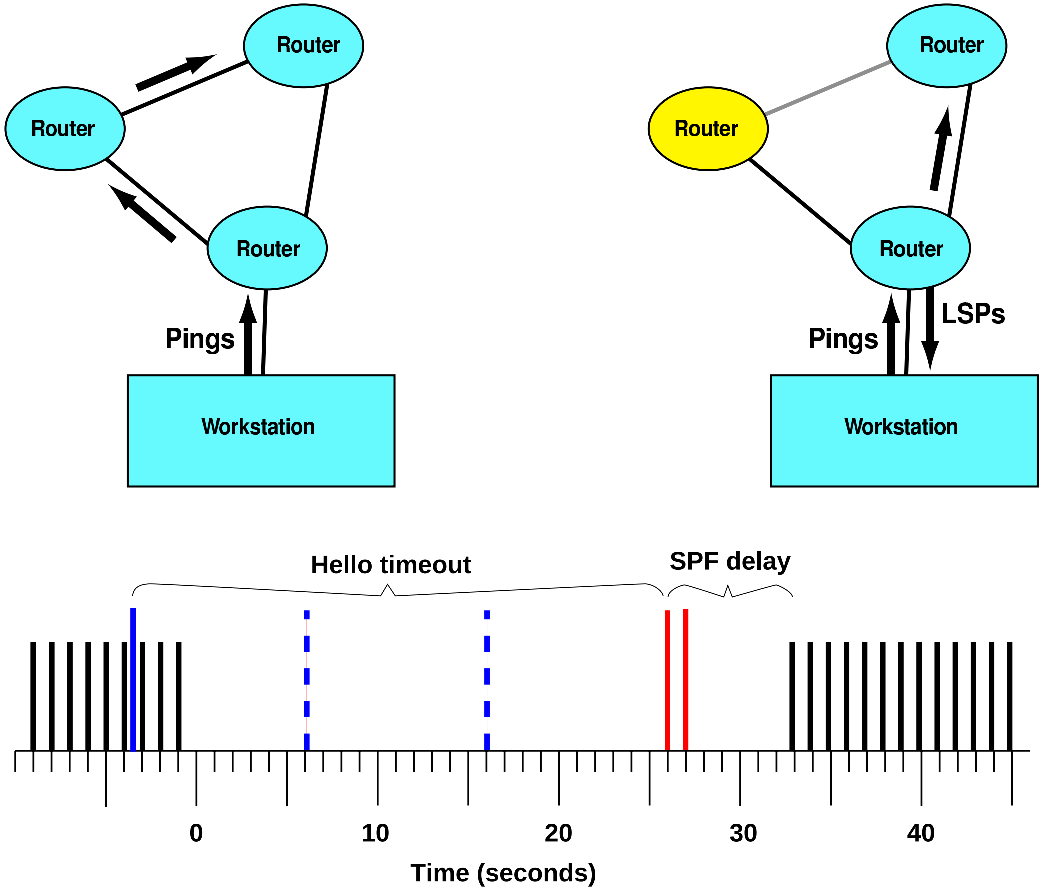
Detection

The IS–IS spec allows for two link state change detection mechanisms: link-level notification and peer–peer Hello packets.

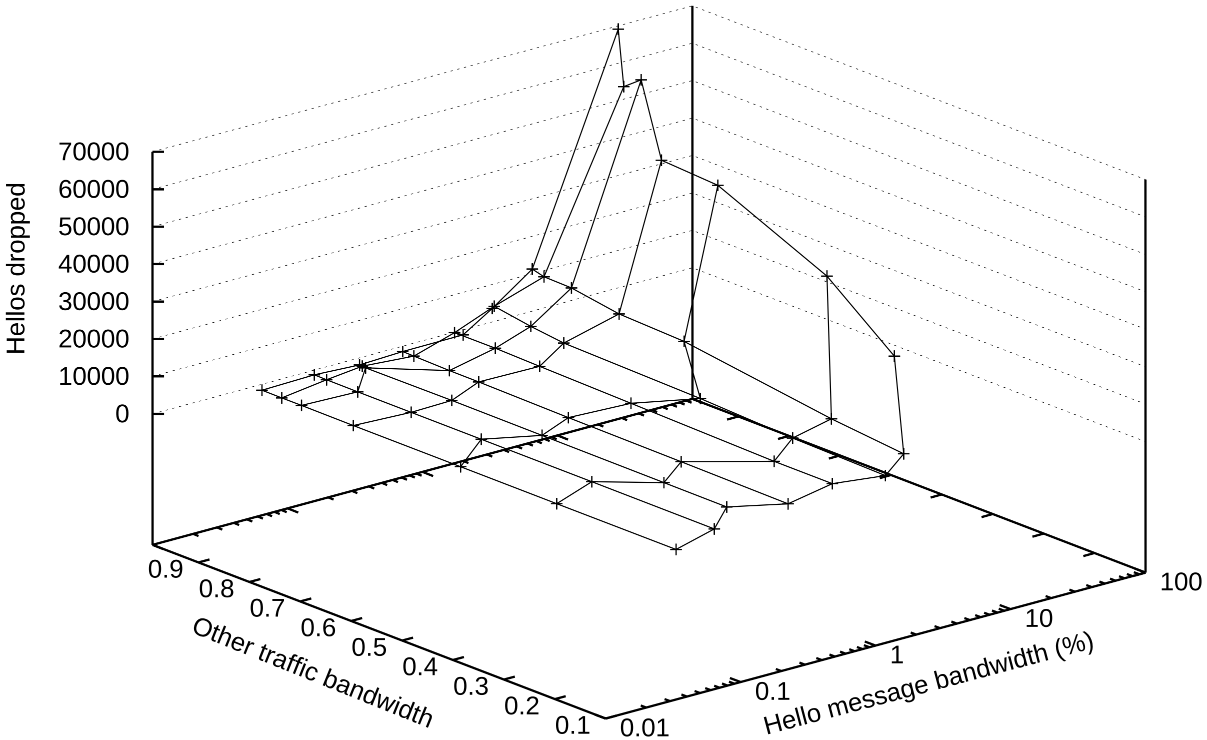
Link level detection should be fastest but it’s not always possible (e.g., switched ethernet) and seems to be inconsistently implemented by vendors.

Spec says that adjacent routers should send Hello packets to each other at a fixed interval (default 10 sec., minimum 1 sec.) and declare the adjacency lost if no Hellos received for three intervals. This works for any interconnect but constrains the repair time to be at least 3 seconds (3 times the Hello interval).

Detection — experiment



Detection — constraints on the Hello interval



Detection — what's possible

Since the *protocol's* ultimate limit on the Hello interval is set by the bandwidth used by Hello packets, extending the spec to allow sub-second intervals would allow sub-second detection on almost all links.

This *doesn't* mean that the detection time can be made arbitrarily small, only that detection should be limited by the physical constraints of links, not by arbitrary clock granularity choices made by protocol designers.

In general, detection time should be limited by the transient error spectrum on a particular link. For example, a link that takes 30ms noise hits should have at least a 30ms Hello interval.

Detection — stability and damping

For either event triggered or hello driven detection, there are network-wide stability issues if routing tries to follow rapid link transients (i.e., a link that goes down and up several times a second).

The usual way of dealing with this is to treat “bad news” differently from “good news” so routing is quick to find an alternate path on any failure but slow to switch back when the link comes up.

The current IS–IS spec treats bad news and good news the same but it should be possible to change the spec to allow different filtering constants for “down” and “up” state changes.

LSP Propagation

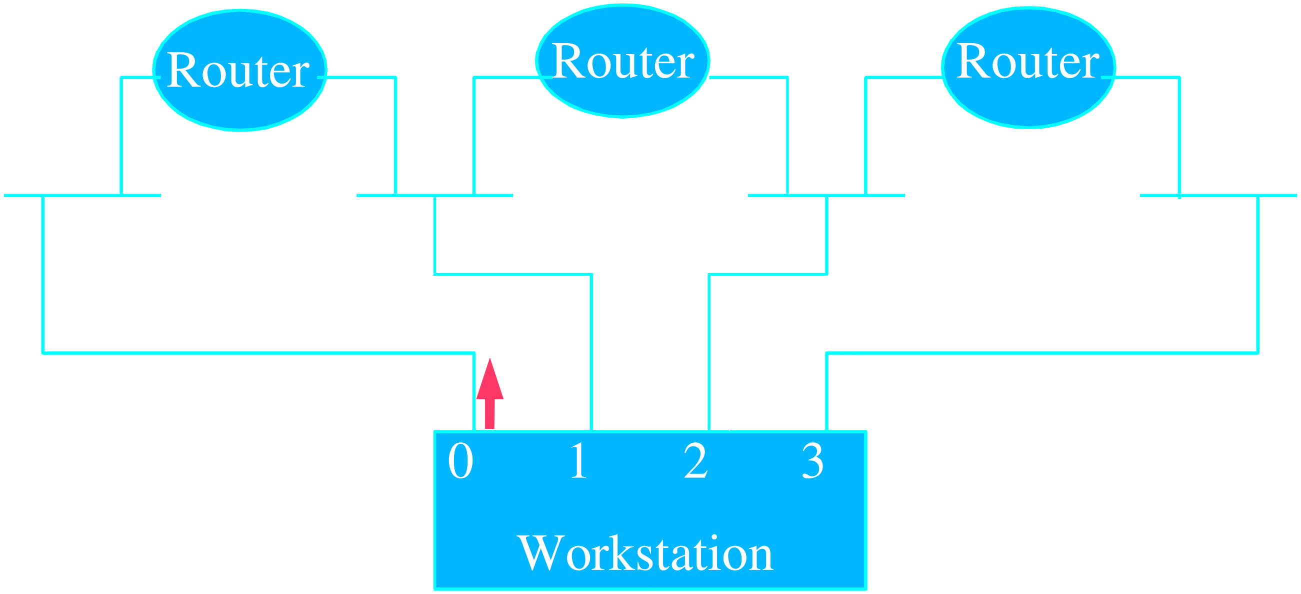
A link state packet is generated at the point of detection then flooded, unmodified, through the network.

It should propagate at near the speed of light plus one store-and-forward delay per hop.

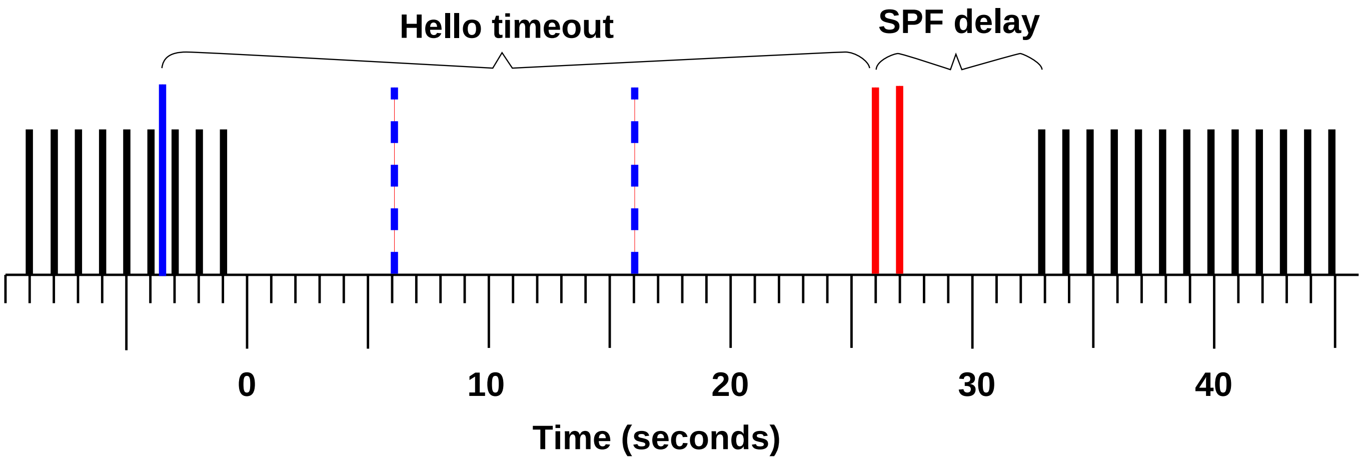
So in theory LSP propagation should make a negligible contribution to the re-route time.

Theory doesn’t often resemble reality...

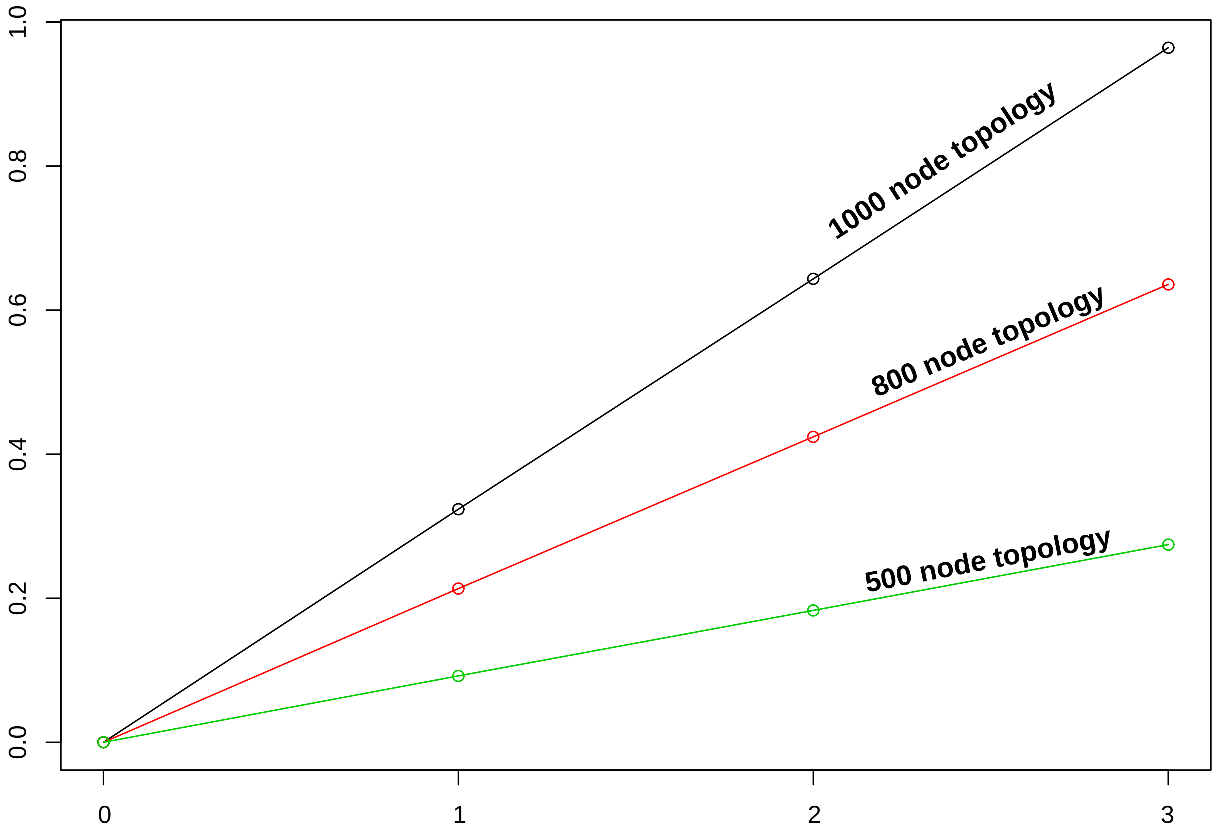
LSP propagation experiment setup



Effect of SPF delay on LSP propagation



LSP propagation experiment with zero SPF delay



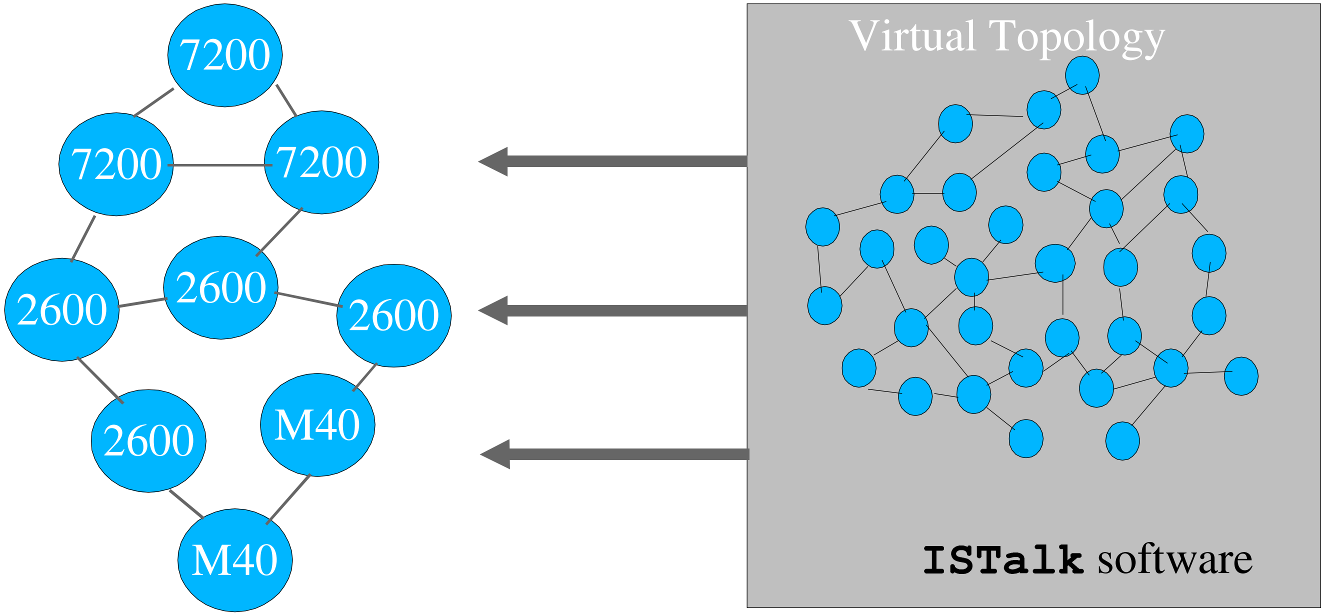
LSP propagation explanation

Since the SPF calculation can take a significant amount of time (see next section), commercial router implementations impose a limit on how frequently the calcularion can be done. In some implementations this limit is fixed (5 seconds) and in some it is changable but with a granularity of 1 second. This limit essentially adds to the propagation time.

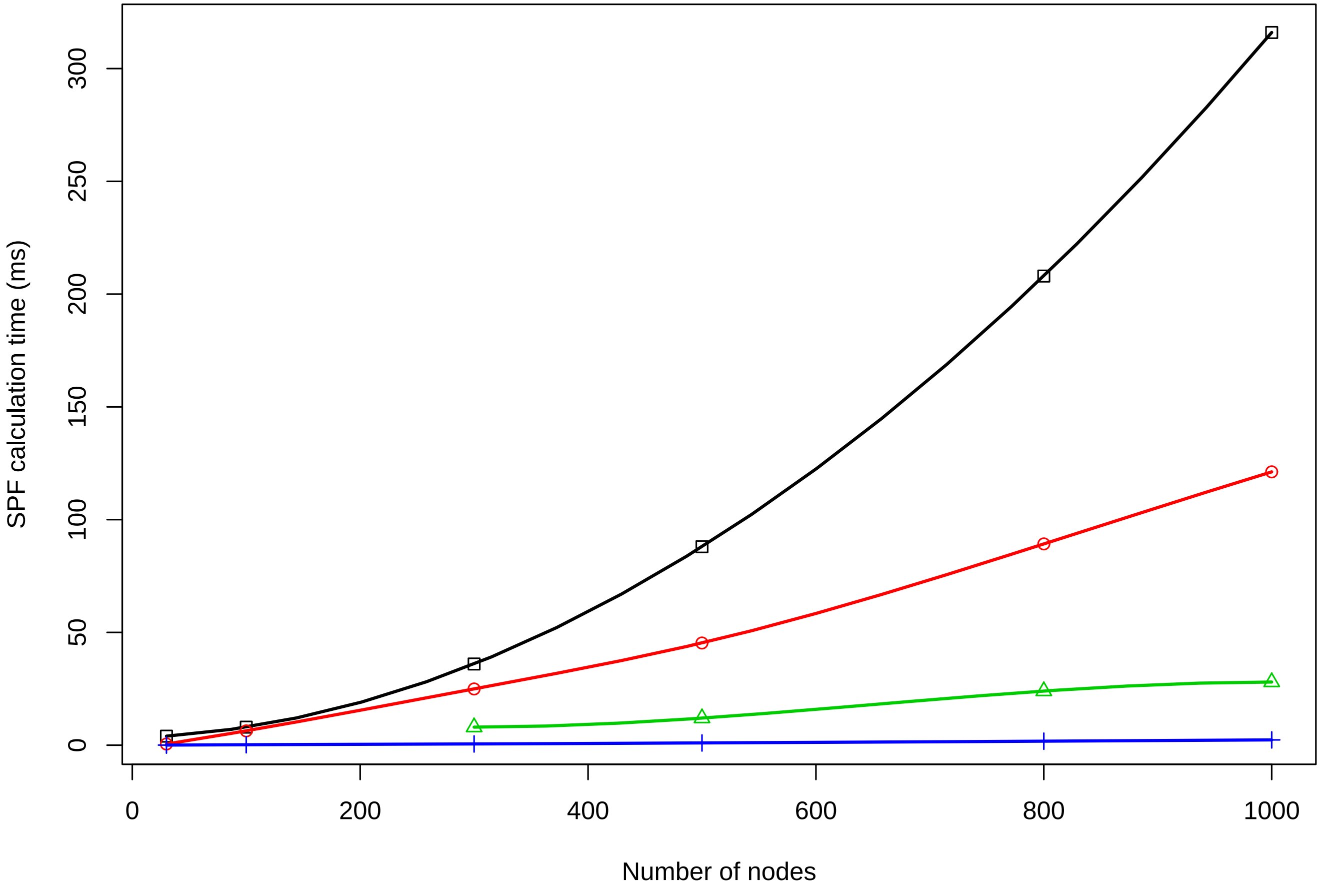
On at least one implementation, if the SPF limit is set to zero the SPF calculation is done before the changed LSPs are flooded which degrades the propagation time from $O(speedoflight)$ to $O(diameter \times SPFtime)$.

To prevent this, the spec might be amended to explicitly state that LSP flooding is “higer priority” than SPF calculation or the point might become moot if the SPF calculation time is improved (next section).

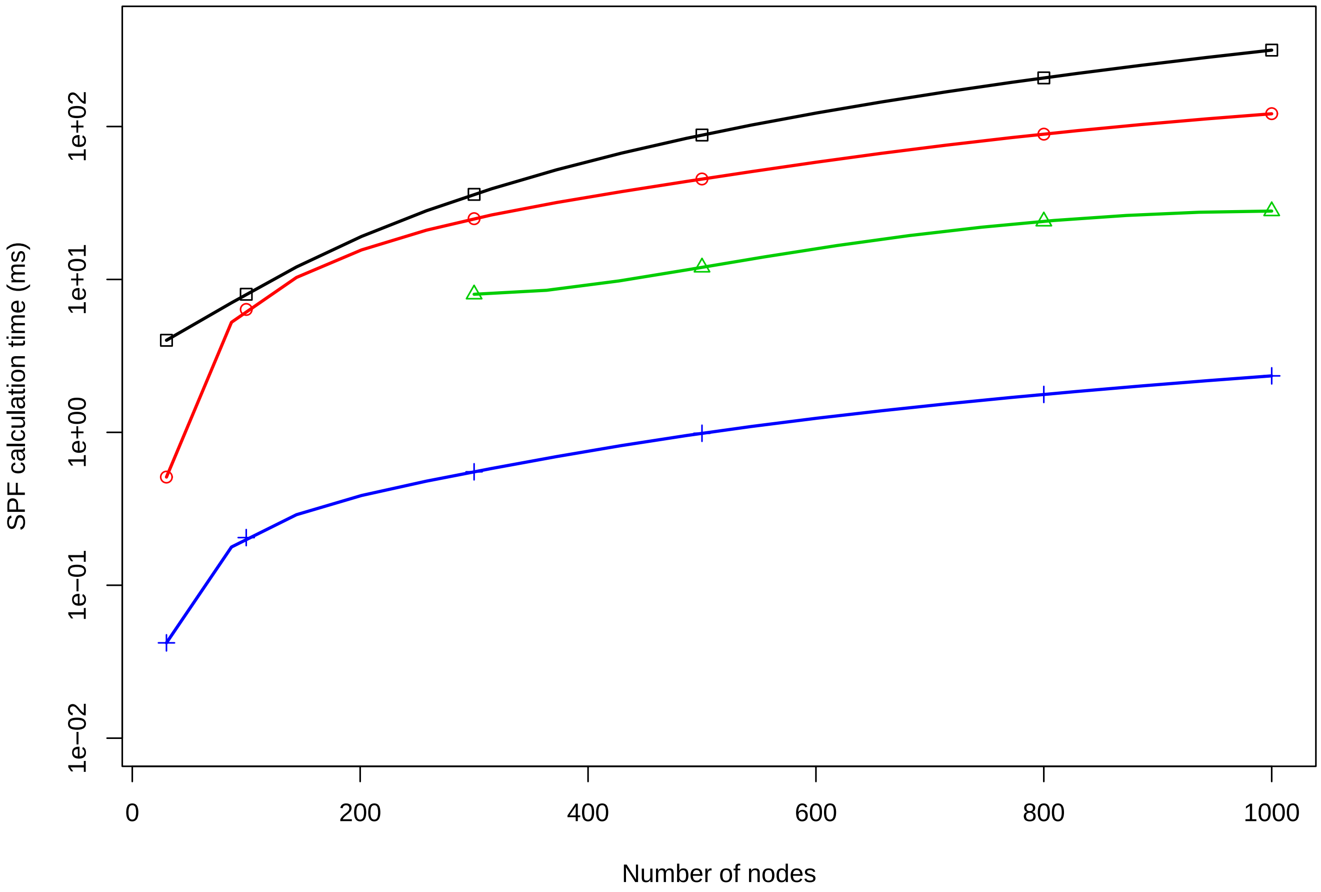
SPF calculation experiment setup



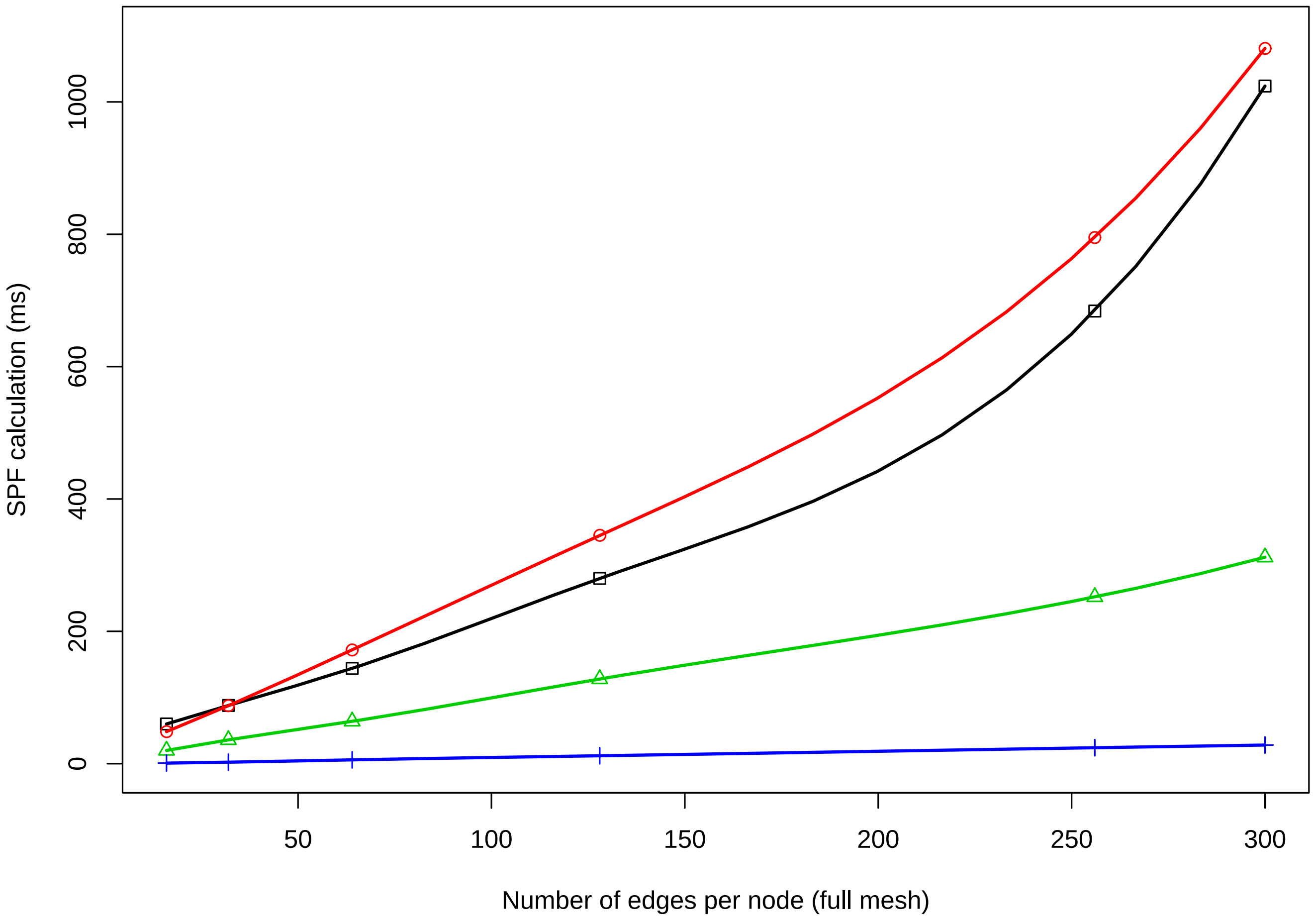
SPF measurements — random topologies



SPF measurements—random topologies (log scale)



SPF measurements — full meshes



Packet Design

SPF calculation discussion

After any link state change, each node has to do an SPF calculation to compute the new topology. Even with high-end platforms (Cisco 7200, Juniper M40) this calculation can take a lot of time (seconds) and has poor scaling properties ($n \log n$ to n^2).

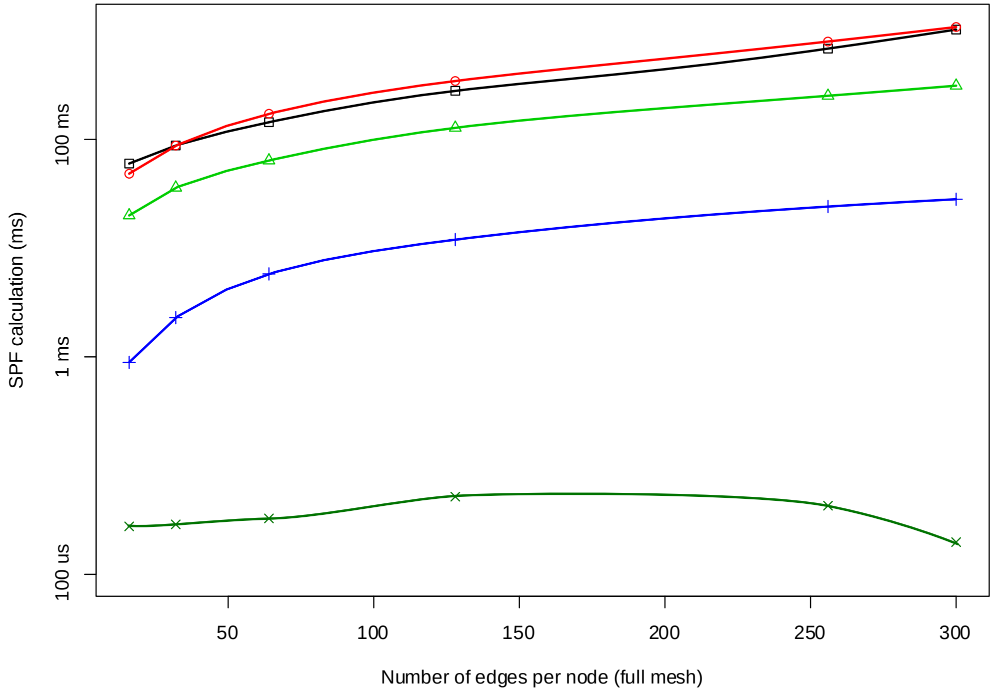
This has a serious impact:

- on convergence (because the SPF calculation is in series with the LSP propagation).

- on overall network stability (because of router CPU saturation).

Fixing the SPF calculation

The Dijkstra SPF algorithm is almost 40 years old. More recent algorithms can compute changes to SPF trees in time proportional to $\log n$ rather than $n \log n$. This allows a net to scale up to virtually any size while bringing the calculation time down from seconds to microseconds.



Final note — what we didn’t see

During our experiments we injected topologies with 50,000+ edges, drove all the routers to 100% CPU saturation, then randomly unplugged links and powered off boxes.

Although we were looking for it, we saw no evidence of instability and we observed several subtle things done expressly to avoid getting into an unstable operating regime.

It appears that the two vendors we looked at have learned a lot from a decade’s worth of routing disasters and meltdowns and are currently shipping some pretty robust routing code.

Summary

Stable, robust IP re-routing that works at the network's propagation rate (the theoretical maximum for *any* re-routing scheme) is both possible and achievable.

To get there, we have to make some minor changes to the IS–IS spec then customers have to let the router vendors know (if) it's worth implementing these changes.

The changes are (in rough priority order):

(1) switch to a modern algorithm for the SPF calculation.

(2) make the granularity of the Hello timer milliseconds rather than seconds.

(3) allow different detection filter constants for link up and down events.

Acknowledgments and a request

The authors are grateful to the Global Crossing Production Testing Lab, Cisco Systems and Juniper Networks for allowing us to use their facilities to perform some of these tests.

We are interested in measuring IS-IS and OSPF behavior on real, operational ISP networks. If you'd like to work with us on this, please contact `cengiz@packetdesign.com`.